doi:10.12118/j.issn.1000-6060.2019.06.20

基于随机森林算法的土壤有机质含量高光谱检测①

包青岭[1,2]， 丁建丽[1,2]， 王敬哲[1,2]， 蔡亮红[1,2]

（1 新疆大学资源与环境科学学院智慧城市与环境建模自治区普通高校重点实验室，新疆 乌鲁木齐 830046；
2 绿洲生态教育部重点实验室，新疆 乌鲁木齐 830046）

摘 要： 为了探讨既能保留光谱信息又能准确对土壤有机质含量进行快速检测。以新疆南部渭干河—库车绿洲内部73个土壤样点及其对应的高光谱数据为研究对象，采用小波变换与数学变换进行光谱数据预处理，分析各小波分解重构光谱在不同有机质含量与不同土壤类型下光谱曲线差异，通过相关分析确定最大小波分解层并筛选敏感波段，结合灰色关联分析与随机森林预测分类模型对各小波分解特征光谱进行重要性分析，最后基于最优特征光谱建立多元线性预测模型并进行分析。结果表明：(1) 耕作土壤与林地土壤光谱曲线波段相较盐渍土壤和荒漠土壤光谱曲线变化较为平缓，同时在水分吸收波段处，盐渍土壤光谱曲线吸收谷最深。(2) 小波变换分解光谱与土壤有机质含量的相关性随着分解层数增加呈现先减后增趋势，在第6层中，特征光谱曲线与敏感波段数量变化趋于稳定，确定为小波变换最大分解层。(3) 随机森林模型相比灰色关联分析对于各小波分解层因子的筛选符合预期，按照对土壤有机质含量影响从高到低排序为 *L*3-(1/Lg*R*)′、*L*4-(1/Lg*R*)′、*L*6-(1/Lg*R*)′、*L*5-(1/Lg*R*)′、*L*2-(1/Lg*R*)′、*L*0-1/Lg*R*、*L*1-1/Lg*R*。(4) 在小波分解光谱中，中频范围特征光谱对干旱区土壤有机质含量的估测能力优于高频与低频范围特征光谱，同时基于 L-MC 建立的模型精度最高。研究表明：基于机器学习分类方法结合小波分解的土壤光谱有机质含量监测，可以有效的减少噪声波段干扰，并提高特征波段的分类预测精度。

关 键 词： 高光谱；土壤有机质含量；小波变换；随机森林

土壤有机质（Soil organic matter，*SOM*）是地球表面土壤中重要组成物质，作为反映土壤肥力以及土地生产能力的重要因子。国内外不同学者利用高光谱技术对不同土壤类型估测 *SOM* 含量，发现 *SOM* 在土壤光谱曲线不存在明显吸收峰，对于可见光至近红外范围存在明显光谱敏感区[1-2]。众多学者利用高光谱数据对 *SOM* 含量进行定量估算，均取得了不错效果[3-9]。由于土壤中存在与 *SOM* 含量不相关的噪声波段，所以有效减少噪声影响并保留光谱有效信息是 *SOM* 光谱定量估算的难点[10]。

现今比较成熟的光谱平滑去噪技术包括 Savitzky-Golay 滤波、中值滤波、移动平均法等。MORGAN 等[11]使用移动加权算法进行土壤有机碳含量的估测，RIENZI 等[12]和 NOCITA 等[13]选用不同采样窗口 Sdavitzky-Gplayl 滤波进行土壤有机碳检测中光谱数据平滑去噪。上述研究虽然能对光谱反射率数据起到去噪和压缩的效果，但是对于白噪声，尤其是随机和低频信号，难以做到去除噪声又不影响有用信号。小波变换方法作为一种新的光谱平滑去噪技术，已经成功应用在高光谱数据处理中。LIAO 等[14]采用4种常用光谱变换方式对 *SOM* 含量进行分析建模，结果表明小波变换对于减少噪声波段方面有很好效果。YANG 等[15]对反射光谱采用小波分析捕捉土壤有机碳和总氮的吸收特征，结果表明连续小波变换对于定位有机碳与总氮光谱特征和区分土壤其它成分是一种高效的方法。LIN 等[16]基于400～1 006 nm 光谱范围内主要吸收谷，通过小波变换放大被噪声掩盖的有用信息，构造偏最小二

① 收稿日期：2019-03-20； 修订日期：2019-07-18
基金项目：新疆自治区重点实验室专项基金（2016D03001）；新疆自治区科技支疆项目（201591101）
作者简介：包青岭（1993-），男，硕士研究生，新疆伊犁人，主要从事土壤遥感研究. E-mail：13139805801@163.com
通讯作者：丁建丽. E-mail：watarid@xiu.edu.cn

乘模型,结果表明小波变换具有很好光谱降噪效果。张锐等[17]研究发现,中频范围小波分解层对 *SOM* 含量的预测较为精确,陈红艳等[18]与王延仓等[19]研究发现高频范围小波分解特征光谱对于 *SOM* 含量的预测较为适合。以上研究表明,相比较传统的光谱去噪方法,小波变换能实现光谱信号的去噪与特征光谱选择[20]。

前人的研究多集中在小波分解特征光谱与 *SOM* 进行定量估算,但是较少考虑通过数据挖掘模型进行小波分解重构光谱结合数学变换的优选并进行 *SOM* 含量的预测,因此本文选择渭干河—库车河三角洲(简称渭—库绿洲)为研究区,选取表层 *SOM* 含量与相对应的土壤光谱进行定量分析并进行建模反演。分析不同 *SOM* 含量下与不同土地利用方式下土壤光谱反射率在各波段与各分解层的差异,根据各层特征光谱曲线与 *SOM* 含量之间相关性确定最佳分解层,并对原始土壤光谱数据和重构光谱分别进行 9 种数学变换,利用随机森林数据挖掘模型与灰色关联分析方法,对小波分解特征光谱因子进行重要性分类,最后进行干旱区 *SOM* 含量的多元线性建模预测,为干旱区土壤养分的研究与当地精准农业提供科学参考与支持。

1 研究区概况与研究方法

1.1 研究区概况

以新疆维吾尔自治区的渭—库绿洲内部区域(41°06′~41°38′N、81°26′~83°17′E)为实验区,其中包括库车、沙雅、新和 3 个县。渭—库绿洲位于新疆塔里木盆地中北部,是新疆具有代表性的干旱区绿洲。年内平均气温范围在 0.5~14.4 ℃,年内平均降水在 67.5 mm 左右。属于典型的极端干旱区域。土壤类型主要以潮土、灌淤土和棕漠土等为主,同时也有水稻土、盐土、草甸土、沼泽土等[21]。依据野外实地采样单元确定研究区范围,如图 1 所示(由 Landsat-OLL 8 影像红绿蓝波段合成)。

1.2 研究方法

1.2.1 土样采集与光谱处理 在 2017 年 7 月中旬,采集了分布于绿洲区域内部的 73 个土壤样品(图 1),覆盖了绿洲内部区域的不同土地利用方式,包括农田、荒地、盐渍地和林地。按照 5 点梅花状采集表层(0~10 cm)土样且将 5 个土样进行混合,将土样带回实验室,进行自然风干、研磨并将土样中的杂质进行剔除后过筛。*SOM* 含量的测定采用重铬酸钾容量—外加热法(油浴)测定。土壤反射率数据的测定采用 ASD Field spec3 便携式光谱仪在暗室内进行,将过 0.25 mm 筛的土壤样品装载进深 1.8 cm、直径 12 cm 盛样皿内,波谱范围在(350~2 500 nm),室内光源为 50 W 的卤素灯,采用 5°视场角光纤探头。光谱测定前均进行白板校正,每个土样测量 10 次,算术平均以后得到土样的实际反射光谱曲线,去除边缘波段(350~399 nm)和(2 401~2 500 nm)。为了消除光谱数据受实验环境、光谱高频随机噪声、杂散光等干扰影响,采用 Savitaky-Golay(2 次多项式,5 个点)平滑去噪[22-23]。

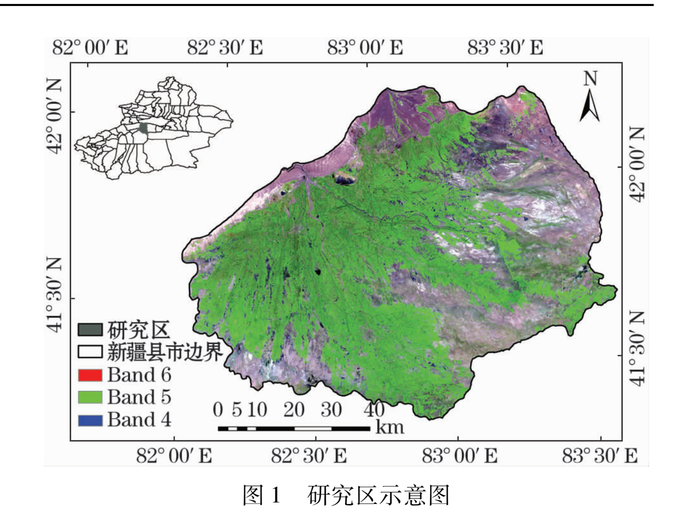


图 1 研究区示意图

Fig. 1 Map of the study area

1.2.2 小波分解 小波分析是一种基于傅里叶变换法发展起来的数据分析方法,小波多尺度分解是通过构造小波基函数对分析函数进行多尺度分解,常见的小波变换有连续小波变换(Continuous wavelet transform, CWT)和离散小波变换(Discrete wavelet transform, DWT)。小波分解将原始光谱信号分解为不同子频带的时频分量,从而更好地观察原始信号的特定频率特征,小波分解的每一层子频带可表示为原始光谱某一频率的吸收特征,相对应的高频光谱信号则被小波滤波器去除[24-25]。根据王延仓等研究结论,因此本研究选取 db5 小波母函数对原始光谱进行 8 层小波变换,并构建各层特征光谱,以 *L1*~*L8* 表征,最后再选取与 *SOM* 含量相关性较好的 *L1*~*L6* 层特征光谱进行进一步的分析。

1.2.3 随机森林模型 随机森林模型(Random Forest Model, RFM)属于机器学习的一大分支—集成学习(Ensemble Learning)方法。RFM 算法是基于决策树分类集成算法,其中每一棵树都依赖于一个

随机向量,通过对数据集的列变量和行变量观测进行随机化,生成多个分类数,最终将分类树结果进行汇总。RFM 对于非线性问题有很好的解释能力,相比于神经网络,降低了运算量的同时也提高了预测精度。本文在 R 语言中,利用 Random Forest 工具包进行预测分类,在进行拟合前,分别对需要生成树的数量(B = ntree)参数设定为 600,每个节点处用于分割节点的预测变量树(d = mtry)参数设定为 3。模型的重要性分类指标由平均下降精度参数(Mean Decrease Accuracy)提供,模型的预测性能可以通过预测相关系数(R^2)、均方根误差($RMSE$)2 个指标来衡量 RFM 预测性能。RFM 的 R^2 越大,$RMSE$ 越小,其 RFM 估算准确性越高,反之则准确性越差[26]。

1.2.4 ***数据预处理*** 在确定最大小波分解尺度的基础上,将经过小波分解的各层光谱特征数据进行 9 种常规数学变换,这 9 种数学变换包括对数(LgR)、倒数(1/R)、倒数的对数(Lg1/R)、对数的倒数(1/LgR)、一阶微分(R')、倒数的对数的一阶微分〔(Lg1/R)′〕、对数的倒数的一阶微分〔(1/LgR)′〕、对数的一阶微分〔(LgR)′〕、倒数的一阶微分〔(1/R)′〕。这些数学变换在 Excel 和 Oringin9.2 中进行,小波分解在 MatlabR2012a 进行操作[27]。

1.2.5 ***数据分析与建模验证*** 选取原始光谱与重构光谱与 SOM 相关性最大的波段为多元逐步回归模型的自变量,SOM 含量为模型的因变量。并且参照 SOM 含量与重构光谱随机森林分类结果对应的特征波段作为多元逐步回归建模自变量 L-MC,SOM 含量作为模型因变量。模型精度评价参数有:校正决定系数(Determination of cofficients,R_c^2)、验证决定系数(Determination coefficients of validation,R_p^2)、残留预测偏差(Residual prediction deviation,RPD),其中当 $RPD \geq 2$ 时,模型达到精准;当 $1.4 \leq RPD \leq 2$ 时,模型精度可靠;当 $RPD < 1.4$ 时,模型并不可靠[28]。

2 结果

2.1 研究区土壤有机质含量描述

SOM 含量的基本描述情况如表 1 所示。可知研究区所有土样集的 SOM 含量平均值为 32.93 g·kg^{-1},校正集与验证集对应的有机质含量分别为 30.63 g·kg^{-1}和 41.21 g·kg^{-1}。全样本集、校正集和验证集的变异系数(Coefficient of variation,CV)分别为 40.57%、39.35% 和 29.53%,属于中等变异。由表 2 可知,不同土地利用类型中,林地 SOM 含量均值最大,为 47.80 g·kg^{-1},依次为农田、盐渍地与荒地。标准差最大值与最小值分别为林地 14.05 g·kg^{-1}与荒地 8.59 g·kg^{-1},各地类变异系数属于中等变异。

2.2 不同 SOM 含量小波变换分析

选取 SOM 含量差异较大的 3 种土样,分别为 41.32 g·kg^{-1}、33.91 g·kg^{-1}和 22.71 g·kg^{-1},探究各土样小波分解特征光谱之间的差异。

同 SOM 含量下小波变换后重构光谱如图 2 所示,从不同含量 SOM 经小波分解后的室内光谱曲线图 2 可以看出,不同 SOM 含量光谱曲线在 $L1$ ~ $L8$ 小波分解层中形态较为一致,整体上呈现上凸的抛物线形。根据以往研究[29],当 SOM 含量大于 2% 时,SOM 含量则在描述土壤光谱反射率特性中起主要作用。在 $L0$ ~ $L8$ 小波分解重构光谱中,每一层光谱反射率都是随着 SOM 含量增加而降低。在400 ~ 800 nm 之间,每一层重构光谱都形成一个陡坡,反射率在此波段范围内增加较快,同时随着分解层数的增加,光谱曲线逐渐变得平滑,消除了大部分噪声,直到 $L7$ 层,光谱曲线逐渐变成一条直线。在近红外区域,反射率变化较为平缓,同时形成以 1 400 nm、1 900 nm、2 200 nm 波段为主的水分吸收谷,随着分解层数的增加,水分吸收谷逐渐变得平坦,直到 $L8$ 层 33.91 g·kg^{-1}和 22.71 g·kg^{-1}有机

表 1 SOM 含量描述性统计

Tab.1 Statistical characteristics of SOM of soil samples

样品集	土样数	均值 / g·kg^{-1}	标准差 / g·kg^{-1}	最小值 / g·kg^{-1}	最大值 / g·kg^{-1}	变异系数 / g·kg^{-1}
全样本集	73	32.93	15.16	15.13	70.77	40.57
校正集	51	30.63	14.56	15.13	70.77	39.53
验证集	22	41.21	13.79	13.04	69.66	29.53

表 2 土壤表层不同土地利用方式下 SOM 含量

Tab.2 SOM contents in top soil under different type of land use

类型	土样数	最小值 / g·kg^{-1}	最大值 / g·kg^{-1}	均值 / g·kg^{-1}	标准差 / g·kg^{-1}	变异系数 / %
农 田	31	18.55	70.77	35.01	13.40	38.29
林 地	5	27.85	69.66	47.80	14.05	29.41
盐渍地	30	15.13	53.53	32.13	10.43	32.46
荒 地	6	13.04	37.21	29.05	8.59	29.57

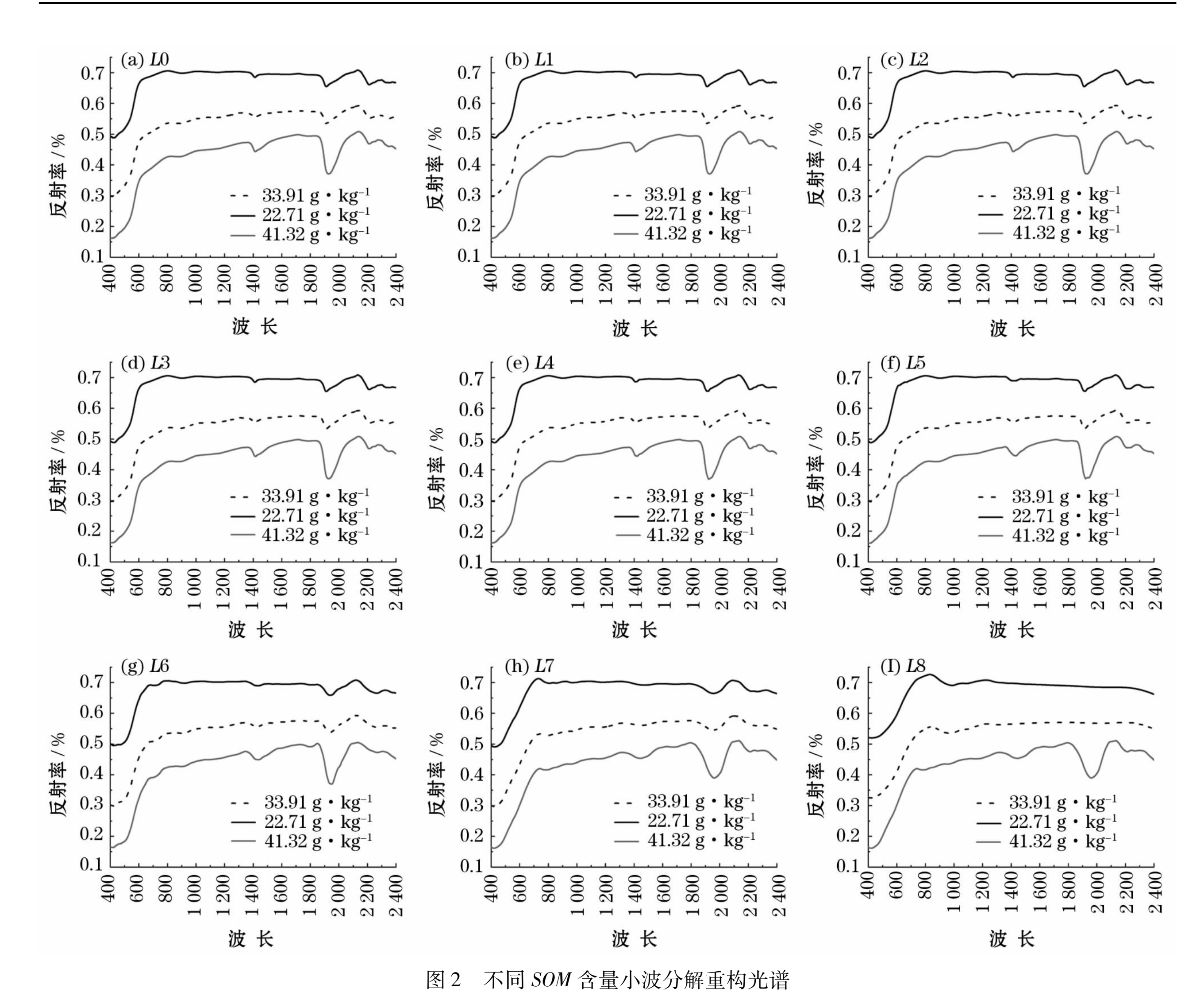


图 2　不同 *SOM* 含量小波分解重构光谱

Fig 2　Wavelet decomposition and reconstruction spectra under different *SOM* contents

质含量的光谱曲线已经看不到水分吸收谷,在 850 nm 波段范围内有明显吸收。

2.3　不同土地利用方式下小波变换分析

图 3 为 4 种不同土地利用方式下经小波变换后的土壤光谱曲线,分别为耕作土壤、林地土壤、盐渍土壤和荒漠土壤,同时 4 种不同类型土壤的 *SOM* 含量均在 22.63 g · kg^{-1}附近。通过 *L0* 原始光谱曲线发现,在 400 ~ 800 nm 范围内,4 种土壤类型光谱曲线随着波长增加,反射率急剧上升,形成 4 个反射峰,荒漠土壤反射率上升最快,曲线斜率最大,直到 900 nm 左右超越其它类型土壤,之后保持反射率第一的位置,荒漠土壤与盐渍土壤的光谱曲线在500 ~ 900 nm 之间存在一个明显的弓形突起区,该发现与高志海等[29]研究相符。在以 1 400 nm、1 900 nm 和 2 200nm 波段为主的水分吸收谷,盐渍土壤的水分吸收谷最深,依次为荒漠土壤、耕作土壤与林地土壤,同时随着小波分解的进行到 *L8* 层,只有盐渍土壤光谱曲线还有着明显水分吸收谷。在可见光范围内,这 4 种土壤类型光谱曲线出现交叉现象,这与李洪等[30]的研究较为一致。在 1 000 ~ 2 400 nm 波段范围内,林地土壤光谱曲线变化较为平衡,盐渍土壤光谱曲线波动最为剧烈,接下来为荒漠土壤与耕作土壤。经过小波变换后,在 *L4* 与 *L5* 层去噪效果达到最佳,光谱曲线基本无毛躁现象,同时又很好的保留了不同土壤类型光谱曲线的特征,但是在 *L7* 与 *L8* 层中,已经消除了大部分光谱特征。

2.4　SOM 与重构光谱相关性分析

对每个土壤样本反射率进行小波变换,变换后得到各层分解特征光谱,并且在此基础上进行 9 种常规数学变换,最后得到土壤反射光谱的敏感波段。选取 8 层分解光谱反射率与 *SOM* 含量进行相关性分析,表 3 与表 4 中的相关系数均通过 0.01 置信水

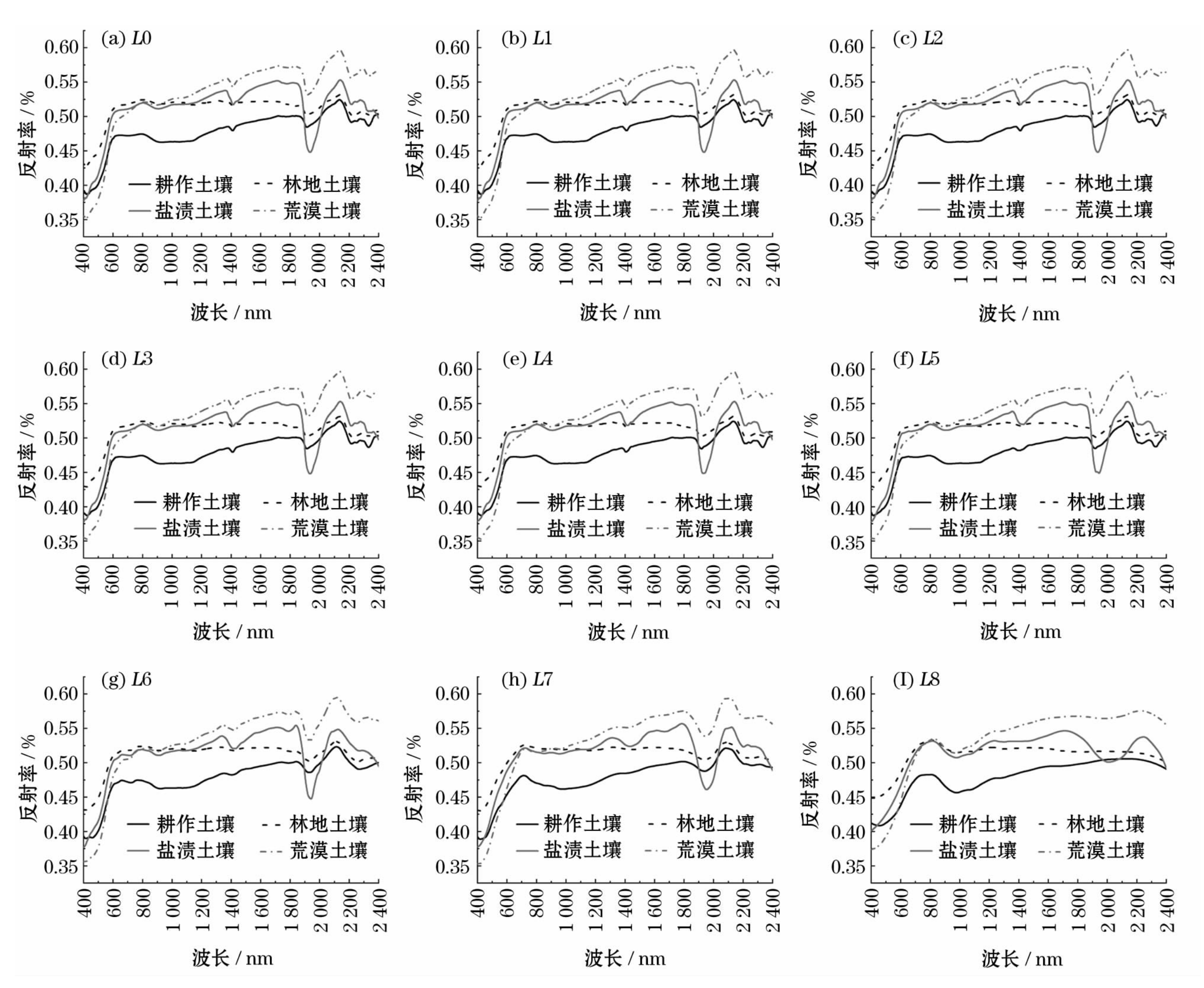


图 3　不同土地利用方式下小波分解重构光谱

Fig 3　Wavelet decomposition and reconstruction spectra under different land use type

平下的 F 检验。在表 3 中，$L1$ ~ $L3$ 层通过相关性显著性检验波段的数量几乎一致，同时敏感波段基本集中在 2 337 nm 附近，最大相关系数为 0.435 8，变化趋势不明显，随着分解层数的增加，*SOM* 的显著性波段在 $L6$ 达到最多，但是在 $L7$ 与 $L8$ 分解光谱的显著性波段数快速减少，同时最大相关系数也迅速减少。由于 $L6$ 层分解光谱不仅能去噪，还能最大程度保留光谱信息，因此本研究选择最大分解层数为 6 层，并在 $L1$ ~ $L6$ 的基础上进一步分析。

表 3　*SOM* 与各层特征光谱相关分析

Tab. 3　Correlation analysis between *SOM* and spectra from wavelet analysis in each level

小波分解层	敏感波段数	波段	最大相关系数
$L1$	302	2 338	0.435 8
$L2$	298	2 337	0.435 6
$L3$	299	2 337	0.435 5
$L4$	299	2 337	0.434 7
$L5$	304	2 330	0.431 4
$L6$	268	2 310	0.425 3
$L7$	293	2 320	0.416 7
$L8$	285	2 348	0.414 0

以光谱反射率 R 及其 9 种常规数学变换与 *SOM* 含量的相关系数通过显著性检验的最大相关系数以及所在波段位置进行统计，如表 4 所示。观察发现前 4 种数学变换敏感波段基本在 2 300 ~ 2 400 nm之间，且最大相关系数在 0.40 ~ 0.50 左右，后 4 种经过微分变换的敏感波段集中在 2 100 ~ 2 300 nm 之间，且最大相关系数在 0.45 ~ 0.70 之间。此外，由表 4 看出，在经过 $(1/\mathrm{Lg}R)'$ 数学变换，各敏感波段在此处均出现了较高的相关系数，其他微分处理均很好的提高了相关性，观察得出各分解层数经过微分变换后均极大提高了与 *SOM* 含量的相关性，同时最大相关系数集中在 $L4$ 重构光谱范

表 4　*SOM* 与重构光谱数学变换的最大相关性及其波段所处位置

Tab. 4　Maximum correlation and the location of its band between *SOM* and different mathematical transformation of reconstruction spectra

	R	Lg*R*	1 / *R*	Lg1 / *R*	1 / Lg*R*	*R*′	(Lg*R*)′	(1 / *R*)′	(Lg1 / *R*)′	(1 / Lg*R*)′	*R*′
L0	Band	2 337	2 337	2 337	2 337	2 339	2 110	2 304	844	2 304	2 109
	R^2	0.436	0.417	−0.398	−0.417	−0.484	0.449	0.445	−0.477	−0.446	−0.609
*L*1	Band	2 338	2 338	2 338	2 337	2 338	2 288	2 288	843	2 288	2 109
	R^2	0.436	0.417	−0.398	−0.417	−0.484	0.468	0.456	−0.465	−0.455	−0.620
*L*2	Band	2 337	2 336	2 336	2 337	2 338	2 283	2 283	2 283	2 283	2 110
	R^2	0.436	0.417	−0.398	−0.417	−0.484	0.498	0.496	−0.489	−0.496	−0.596
*L*3	Band	2 337	2 337	2 336	2 337	2 339	2 280	2 283	2 283	2 283	2 280
	R^2	0.436	0.417	−0.398	−0.417	−0.484	0.547	0.526	−0.511	−0.527	−0.655
*L*4	Band	2 337	2 335	2 334	2 335	2 338	2 281	2 281	2 281	2 281	2 281
	R^2	0.435	0.416	−0.397	−0.416	−0.483	0.567	0.532	−0.505	−0.532	−0.688
*L*5	Band	2 330	2 328	2 327	2 328	2 335	2 212	2 210	2 210	2 210	2 276
	R^2	0.431	0.413	−0.394	−0.413	−0.479	0.528	0.535	−0.524	−0.534	−0.635
*L*6	Band	2 310	2 309	2 308	2 309	2 325	2 273	2 216	2 216	2 216	2 273
	R^2	0.425	0.406	−0.388	−0.406	−0.475	0.461	0.479	−0.524	−0.478	−0.527

注:Band 代表最大相关系数波段的所在位置,*R* 代表最大相关系数,*L*0 代表没有经过小波变换的原始光谱

围的 2 281 nm 波段周围。

2.5　*SOM* 与重构光谱重要性分析

以 *SOM* 含量为因变量,在 *L*0 至 *L*6 小波分解光谱中,选择每一层及其 9 种数学变换中相关系数最大的波段的反射率共 7 种因子作为模型的自变量,建立随机森林分类预测模型,表 5 为 *RF* 模型精度拟合结果。观察得,训练集的 R^2 为 0.68,*RMSE* 为 2.11,测试集 R^2 为 0.70,*RMSE* 为 2.45。

图 4 列出了 7 种因子对 *SOM* 含量的影响贡献度, 即 *L*0-1/Lg*R*、*L*1-1/Lg*R*、*L*2-(1/Lg*R*)′、*L*3-(1/Lg*R*)′、*L*4-(1/Lg*R*)′、*L*5-(1/Lg*R*)′、*L*6-(1/Lg*R*)′,同时 Mean Decrease Accuracy 分别为 17.41、12.97、8.04、6.82、6.16、2.87、2.74,并按照从高到低进行排序。由图 4 观察得,对 *SOM* 含量影响较大的因子为 *L*3-(1/Lg*R*)′,其次为 *L*4-(1/Lg*R*)′、*L*6-(1/Lg*R*)′、*L*5-(1/Lg*R*)′、*L*2-(1/Lg*R*)′、*L*0-1/Lg*R*、*L*1-1/Lg*R*。小波分解光谱中频范围,即 *L*3 与 *L*4 层结合(1/Lg*R*)′数学变换对 *SOM* 预测贡献度最大,高频与低频范围,即 *L*2、*L*5、*L*6 层结合(1/Lg*R*)′

表 5　*SOM* 含量随机森林模型模拟精度

Tab. 5　Simulation accuracy of random organic forest model of soil organic matter content

	R^2	*RMSE*
训练集	0.68	2.11
测试集	0.70	2.45

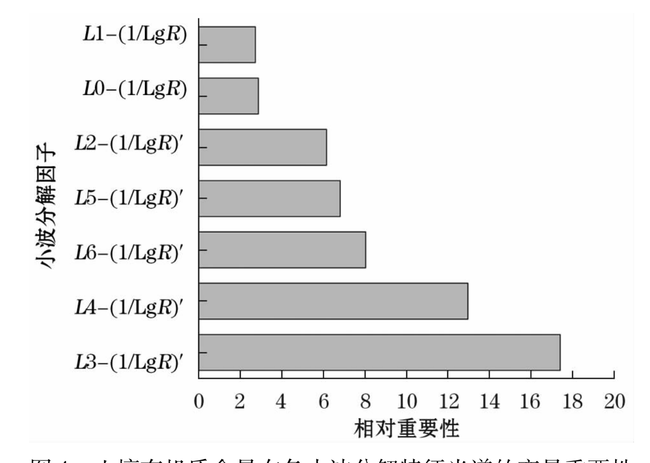


图 4　土壤有机质含量在各小波分解特征光谱的变量重要性

Fig. 4　Importance of soil organic matter content in the spectral characteristics of each wavelet decomposition

数学变换对 *SOM* 含量影响较小,*L*0-1/Lg*R* 与 *L*1-1/Lg*R* 因子对 *SOM* 含量预测贡献度最小。

利用灰色关联分析法对数学变换后的 6 层重构光谱与 *SOM* 含量进行分析,并对其结果进行排序,见表 6。除了 *L*0 原始光谱以及不同数学变换与 *SOM* 含量的灰色关联度,其它重构光谱灰色关联度排序大概相似。根据 *L*1,其排列顺序为 *R* > 1/Lg*R* > 1/*R* > (1/Lg*R*)′ > Lg1/*R* > Lg*R* > (1/*R*)′ > (Lg*R*)′ > (Lg1/*R*)′ > *R*′。通过对比灰色关联分析与随机森林建模分类,研究发现在灰色关联分析下,各层重构光谱关联度排序为原始光谱第一,其次是各数学变换重构光谱,并且无法区分各层重构光谱的纵向排序,相反随机森林分类方法,各因子在重要

表 6　各层重构光谱与不同数学变换的灰色关联分析

Fig. 6　Gray relational analysis of different mathematical transformation of reconstruction spectra of each level

		R	Lg*R*	1/*R*	Lg1/*R*	1/Lg*R*	*R*′	(Lg*R*)′	(1/*R*)′	(Lg1/*R*)′	(1/Lg*R*)′
L0	关联度(*P*=0.5)	0.5662	0.5449	0.56	0.5449	0.5542	0.5863	0.4341	0.4439	0.4335	0.5537
	排序	2	7	3	6	4	1	9	8	10	5
L1	关联度(*P*=0.5)	0.9084	0.894	0.8993	0.894	0.9064	0.3336	0.4006	0.8363	0.4003	0.8945
	排序	1	6	3	5	2	10	8	7	9	4
L2	关联度(*P*=0.5)	0.6673	0.6428	0.6562	0.6428	0.6579	0.4496	0.4441	0.4301	0.444	0.6521
	排序	1	5	3	6	2	7	8	10	9	4
L3	关联度(*P*=0.5)	0.7852	0.7619	0.772	0.7619	0.7794	0.4926	0.5708	0.5736	0.571	0.4699
	排序	1	5	3	4	2	9	8	6	7	10
L4	关联度(*P*=0.5)	0.8979	0.8825	0.8883	0.8825	0.8956	0.6016	0.6255	0.6414	0.6253	0.5029
	排序	1	4	3	5	2	9	7	6	8	10
L5	关联度(*P*=0.5)	0.7215	0.6972	0.7094	0.6972	0.7134	0.4622	0.6651	0.6468	0.6653	0.3757
	排序	1	4	3	5	2	9	7	8	6	10
L6	关联度(*P*=0.5)	0.7924	0.7698	0.7803	0.7698	0.786	0.4012	0.7673	0.7581	0.7673	0.4795
	排序	1	5	3	4	2	10	7	8	6	9

注:关联度在灰色关联分析中代表灰色关联系数,排序在灰色关联分析中代表关联程度

性排序过程中,纵向与横向排列中都能进行有效的区分。研究表明:机器学习分类方法相比较传统线性数学模型分类,具有较好的灵活性、实用性。

2.6　*SOM* 含量建模验证与分析

选取原始光谱与重构光谱中与 *SOM* 相关性最大的波段为多元逐步回归模型的自变量,*SOM* 含量为模型的因变量。并且参照 *SOM* 含量与重构光谱随机森林分类结果,即 *L0* -1/Lg*R*、*L1*-1/Lg*R*、*L2*-(1/Lg*R*)′、*L3*-(1/Lg*R*)′、*L4*-(1/Lg*R*)′、*L5*-(1/Lg*R*)′、*L6*-(1/Lg*R*)′——对应的特征波段作为多元逐步回归建模自变量 *L-MC*,*SOM* 含量作为模型因变量,模型参数和精度参数如表 7 所示。根据表 7 所示的具体每一层重构光谱模型的精度参数进行分析发现,上述 9 个模型中,无论是建模集还是验证集,*RPD*≥1.4 的模型达到 8 个。在 *L0*～*L6* 中,除了 *L6* 外,其它各层重构光谱均能很好的提升模型的精度,同时 *L-MC* 模型,精度最高,R_c^2=0.73,建模 *RPD* 为 1.94,R_p^2=0.74,验证 *RPD* 为 1.96。同时发现基于 *L4* 与 *L3* 重构光谱所构建的模型,预测精度较高,验证 *RPD* 达到 1.80 以上。说明经过小波变换后 *L3*、*L4* 层重构光谱可以一定程度增强光谱对 *SOM* 含量的敏感程度,这 3 种模型对于研究区的 *SOM* 含量具有较好的定量反演能力。所以确定这 3 种模型为最优反演模型。

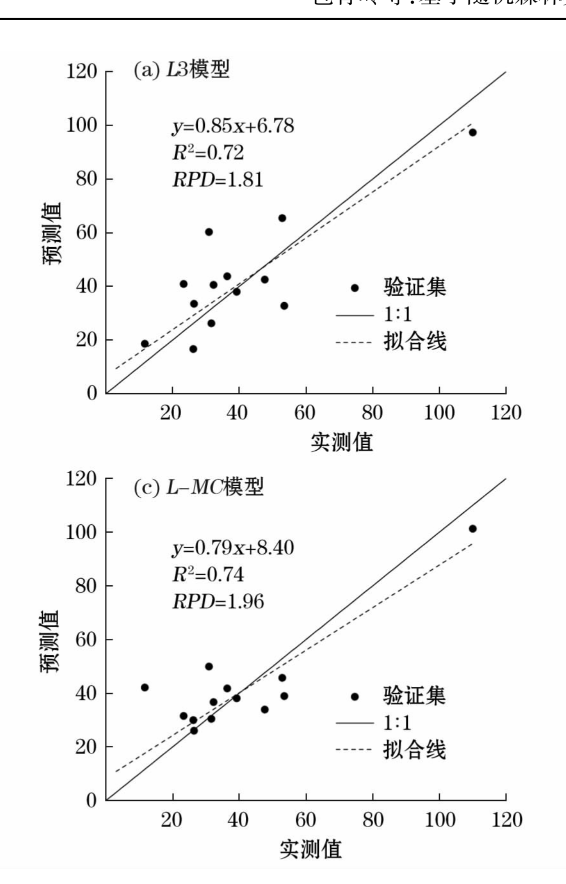
图 5 基于 *L3*、*L4* 和 *L-MC* 模型中实测值与预测值的散点图。由图中可以看出 *L3*、*L4* 和 *L-MC* 中样点基本分布于 1∶1 线附近。各拟合线中,*L-MC* 的系数最小,其中 R^2 达到 0.74,*RPD*=1.96。

表 7　土壤有机质含量反演模型及精度验证

Tab. 7　Inversion models of soil organic matter content and precision validation

变量	模型	建模集		验证集	
		R_c^2	*RPD*	R_p^2	*RPD*
L0	$Y=23.6-6\ 965.58X_{2109J}+54\ 553.28X_{2304G}$	0.68	1.77	0.65	1.60
L1	$Y=14.67-7\ 570.08X_{2109J}$	0.61	1.6	0.62	1.58
L2	$Y=19.48-7\ 371.65X_{2110J}+55\ 148.13X_{2283F}$	0.66	1.72	0.66	1.65
L3	$Y=36.14-42\ 583.03X_{2280J}-206\ 784.62X_{2280F}$	0.64	1.68	0.72	1.81
L4	$Y=33.18-34\ 335.85X_{2281J}-146\ 964.93X_{2281F}$	0.70	1.82	0.72	1.83
L5	$Y=8.15-19\ 945.83X_{2276J}+477\ 430.17X_{2212F}+64\ 501.91X_{2210H}$	0.76	2.06	0.63	1.60
L6	$Y=34.67-11.62X_{2325E}+282\ 270.09X_{2216G}$	0.51	1.43	0.49	1.40
L-MC	$Y=24.84-13\ 355.03X_{2281J(L4)}-3\ 962.40X_{2109J(L0)}$	0.73	1.94	0.74	1.96

注:*J* 代表 (1/Lg*R*)′,*G* 代表(Lg*R*)′,*F* 代表 *R*′,*H* 代表(1/*R*)′



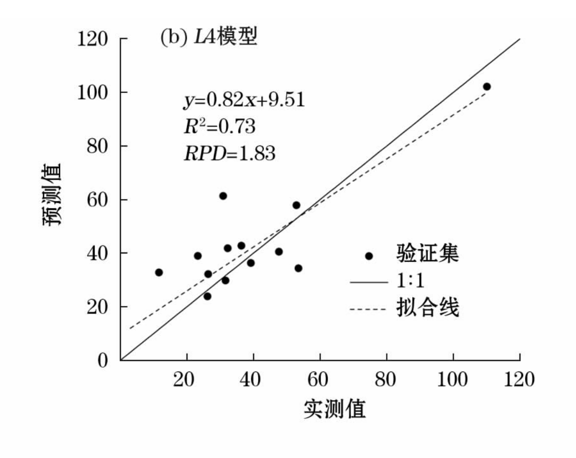


图5 L3、L4与L-MC模型的SOM含量实测值与预测值比较

Fig.5 Measured values of *SOM* in *L3*,*L4* and *L-MC* models were compared with the predicted values

3 讨 论

本研究结果表明,通过小波变换分解光谱结合数学微分变换与随机森林重要性参数分类方法,优选有效的特征波段,将所得结果作为 *SOM* 含量多元线性模型预测的重要因子,可以有效的实现干旱区 *SOM* 含量的快速估测。

研究区为典型的内陆干旱区,绿洲边缘部分盐渍化与荒漠化现象明显,长年累月侵蚀内部农田,通过分析农田土壤、林地土壤与盐渍土壤、荒漠土壤在相同 *SOM* 含量下光谱曲线的差异,发现富含养分的土壤类型与贫瘠土壤类型的光谱曲线在水分波段吸收谷与整体波动存在很大差异,结合小波变换,凸显和简化了不同土壤类型光谱曲线的差异。根据 SHI 等[28]研究土壤光谱曲线反射率随着 *SOM* 含量的升高总体呈下降的趋势,同时 *SOM* 含量与光谱反射率的相关系数较高峰值集中在 400 nm、800 nm、1 400 nm 与 2 200 nm 范围内。如图 2、图 3 和表 3 所示,本研究中,不同 *SOM* 含量下与不同类型土壤的分解特征光谱符合 SHI 等[28]研究,同时 *SOM* 含量与各小波分解特征光谱的相关系数高值集中在 2 200 nm 范围内。高光谱遥感的本质是将待测物连续通道的电磁波谱信息转化为光信号,常用的光谱去噪声手段以微分处理、S-G 平滑、多元散射校正与标准正态变换等方法为主,上述方法在光信号处理上较为适用,但是这些方法在对光谱数据去噪的同时难免会引入新的噪声,而小波变换凭借在时域和频域对于信号的局部化分析能力,通过伸缩平移对信号逐步进行多尺度细化,最终达到高频与低频处信号的细分,在保留原状信号的同时,尽可能的分离噪声,是一种较好的电信号噪声去除方法。结合相关性分析与分解光谱特征分析的结果,本文确定最佳 *SOM* 特征光谱的小波分解层数为 6。陈红艳等[18]研究小波分解后的潮土光谱与 *SOM* 含量的关系,并将小波分解光谱层确定为 9 层;王延仓等[19]研究了小波分解后的北方潮土光谱与 *SOM* 含量的关系,进一步将小波分解光谱层数确定为 6;张锐等[17]研究了小波变换后的原状水稻土光谱与 *SOM* 含量的相关性,相关系数与敏感波段在第 6 层达到最高,以上研究表明,最优分解层不一定都相同,原因包括土壤类型、土壤质地、土壤水分等其他因素。

随机森林预测分类模型相较于传统的线性预测分类模型,比如灰色关联分析,具有显著的优越性。灰色关联分析已经广泛用于土壤光谱研究中,但只能对于土壤的某些特定属性与近红外光谱之间的理想线性关系进行模拟,然而土壤属性与相应近红外

光谱之间的关系不仅仅是简单的线性关系，所以利用简单的线性预测分类，不能有效的反应土壤性质与特征光谱之间的真实关系。随机森林模型在描述两者之间的非线性关系，通常会取得理想的预测分类结果。在两类预测分类模型中，随机森林模型不仅在统计结果上优于灰色关联分析，也在预测能力上表现出更优的可靠性和稳定性。

小波分解将光谱分成了不同频率并重构，低频范围包含了更多高频范围，高频范围更多反映了土壤中全氮、全磷等信息的光谱细节。为了更进一步探讨各小波分解层结合数学变换对 *SOM* 含量光谱预测的影响，利用随机森林预测分类，获得各因子对于 *SOM* 含量预测的重要性，进一步验证了中频范围的小波分解结合(1/Lg*R*)′数学变换不仅能有效去除光谱噪声，还能保持 *SOM* 的光谱细节，解决了土壤光谱有机质信息噪声去除与保留信息之间的矛盾。但是根据王延仓等[19]研究，小波分解重构光谱在低频范围对 *SOM* 含量的估测能力优于高频范围，本实验中，高频范围分解光谱对 *SOM* 含量估测能力较弱，与其相符，但基于中频范围 *L*3 与 *L*4 层模型的估测能力最高，与其不符。可能是土壤类型的不同，导致实验结果不一致。

本文存在一定不足，土壤中其他属性必然会对 *SOM* 光谱预测产生影响，比如土壤水分，如何有效的减少该方面的影响仍需展开研究；随机森林预测分类模型需要更多参数大规模复杂训练，以及对生成树的数量(B = ntree)和预测变量树(d = mtry)参数进行反复设定，选择最优解。本文利用典型样点得到较好的预测精度，下一步研究中，将扩大样点数，对随机森林模型进行训练，获得更可靠的结果，同时将反射光谱与现有的多源遥感系统相结合，为干旱区土壤养分的研究提供科学依据。

4 结 论

本研究以小波变换对原始土壤光谱进行分解重构，分别分析了不同 *SOM* 含量与不同土壤类型的小波分解光谱差异。通过相关性分析和特征光谱分析结果，确定小波分解的最大尺度，对各分解光谱进行数学变换，结合随机森林建模分类与灰色关联分析，分析各因子对于 *SOM* 含量预测的重要性。主要结论如下：

(1) 富含养分的土壤类型较贫瘠土壤类型的光谱曲线在水分波段吸收谷范围内整体波动更大，随着分解层数增加光谱曲线逐渐变得平滑，基本看不出显著差别。

(2) 小波变换不同分解层，从低频到高频范围内与 *SOM* 含量的相关性呈现先减后增的趋势，第 6 层显著波段较多且相关性较高，同时光谱细节保持良好，确定第 6 层为最大分解层数。

(3) 通过灰色关联分析与随机森林预测分类的结果比较，随机森林模型相比灰色关联分析对于各小波分解层因子的筛选符合预期，按照对 *SOM* 含量影响从高到低排序为 *L*3-(1/Lg*R*)′、*L*4-(1/Lg*R*)′、*L*6-(1/Lg*R*)′、*L*5-(1/Lg*R*)′、*L*2-(1/Lg*R*)′、*L*0-1/Lg*R*、*L*1-1/Lg*R*。

(4) 小波分解的中频范围 *L*3 与 *L*4 模型，以及 *L-MC* 模型对干旱区 *SOM* 含量的反射光谱快速估算较为适用。

参考文献(References)

[1] 张勇，庞学勇，包维楷，等. 土壤有机质及其研究方法综述[J]. 世界科技研究与发展，2005. 27(5)：72－78. [ZHANG Yong, PANG Xueyong, BAO Weikai, et al. A review of soil organic matter and its research methods[J]. World Sci-Tech R&D, 2005, 27(5)：72－78.]

[2] 程朋根，吴剑，李大军，等. 土壤有机质高光谱遥感和地统计定量预测[J]. 农业工程学报，2009, 25(3)：142－147. [CHENG Penggen, WU Jian, LI Dajun, et al. Quantitative prediction of soil organic matter content using hyper spectral remote sensing and geostatistics[J]. Transactions of the Chinese Society of Agricultural Engineering, 2009, 25(3)：142－147.]

[3] 何挺，王静，林宗坚，等. 土壤有机质光谱特征研究[J]. 武汉大学学报(信息科学版)，2006, 31(11)：975－979. [HE Ting, WANG Jing, LIN Zongjian, et al. Spectral features of soil organic matter[J]. Geomatics and Information Science of Wuhan University (Information Science Edition), 2006, 31(11)：975－979.]

[4] 刘磊，沈润平，丁国香. 基于高光谱的土壤有机质含量估算研究[J]. 光谱学与光谱分析，2011, 31(3)：762－766. [LIU Lei, SHEN Ruiping, Ding Guoxiang. Studies on the estimation of soil organic matter content based on hyper-spectrum[J]. Spectroscopy and Spectral Analysis, 2011, 31(3)：762－766.]

[5] 叶勤，姜雪芹，李西灿，等. 基于高光谱数据的土壤有机质含量反演模型比较[J]. 农业机械学报，2017, 48(3)：164－172[YE Qin, JIANG Xueqin, LI Xican, et al. Comparison on inversion model of soil organic matter content based on hyperspectral data[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(3)：164－172.]

[6] 郑曼迪，熊黑钢，乔娟峰，等. 基于宽波段与窄波段综合光谱指数的土壤有机质遥感反演[J]. 激光与光电子学进展，2018, 55(7). [ZHENG Mandi, XIONG Heigang, QIAO Juanfeng, et al. Remote sensing inversion of soil organic matter based on broad band and narrow band comprehensive spectral index[J]. Laser & Optoe-

lectronics Progress,2018,55(7).]

[7] 郑曼迪,熊黑钢,乔娟峰,等.基于高光谱的不同人类干扰程度下荒漠土壤有机质含量估算模型[J].干旱区地理,2018,41(2):167－175.[ZHENG Mandi,XIONG Heigang,QIAO Juanfeng,et al. Hyperspectral based estimation model about organic matter in desert soil at different levels of human disturbance[J]. Arid Land Geography,2018,41(2):167－175.]

[8] 彭杰,张杨珠,庞新安,等.新疆南部土壤有机质含量的高光谱特征分析[J].干旱区地理,2010,33(5):740－746.[PENG Jie,ZHANG Yangzhu,PANG Xinan,et al. Hyperspectral features of soil organic matter content in south Xingjiang[J]. Arid Land Geography,2010,33(5):740－746.]

[9] 沈润平,丁国香,魏国栓,等.基于人工神经网络的土壤有机质含量高光谱反演[J].土壤学报,2009,46(3):391－397.[SHEN Ruiping,DING Guoxiang,WEI Guoshuan,et al. Retrieval of soil organic matter content from hyper-spectrum based on ANN[J]. Acta Pedologoca Sinica,2009,46(3):391－397.]

[10] 于雷,洪永胜,朱亚星,等.去除土壤水分对高光谱估算土壤有机质含量的影响[J].光谱学与光谱分析,2017,37(7):2146－2151.[YU Lei,Hong Yongsheng,ZHU Yaxing,et al. Removing the effect of soil moisture content on hyperspectral reflectance for the estimation of soil organic matter content[J]. Spectroscopy and Spectral Analysis,2017,37(7):2146－2151.]

[11] MORGAN C L S,WAISER T H,BROWN D J,et al. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy[J]. Geoderma,2009,151(3－4):249－256.

[12] RIENZI E A,MIJATOVIC B,MUELLER T G,et al. Prediction of soil organic carbon under varying moisture levels using reflectance spectroscopy[J]. Soil Science Society of America Journal,2014,78(3):958－967.

[13] NOCITA M,STEVENS A,NOON C,et al. Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy[J]. Geoderma,2013,199:37－42.

[14] LIAO Q H,GU X H,LI C J,et al. Estimation of fluvo-aquic soil organic matter content from hyperspectral reflectance based on continuous wavelet transformation[J]. Transactions of the Chinese Society of Agricultural Engineering,2012,28(23):132－139.

[15] YANG H F,QIAN Y R,YANG F,et al. Using wavelet transform of hyperspectral reflectance data for extracting spectral features of soil organic carbon and nitrogen[J]. Soil Science,2012,177(11):674－681.

[16] LIN L,WANG Y,TENG J Y,et al. Hyperspectral analysis of soil organic matter in coal mining regions using wavelets,correlations,and partial least squares regression[J]. Environmental Monitoring & Assessment,2016,188(2):97.

[17] 张锐,李兆富,潘剑君.小波包—局部最相关算法提高土壤有机碳含量高光谱预测精度[J].农业工程学报,2017,33(1):175－181.[ZHANG Rui,LI Zhaofu. PAN Jianjun. Coupling discrete wavelet packet transformation and local correlation maximization improving prediction accuracy of soil organic carbon based on hyperspectral reflectance[J]. Transactions of the Chinese Society of Agricultural Engineering,2017,33(1):175－181.]

[18] 陈红艳,赵庚星,李希灿,等.基于小波变换的土壤有机质含量高光谱估测[J].应用生态学报,2011,22(11):2935－2942.[CHEN Hongyan,ZHAO Gengxing,LI Xican,et al. Hyper-spectral estimation of soil organic matter content based on wavelet transformation[J]. Chinese Journal of Applied Ecology,2011,22(11):2935－2942.]

[19] 王延仓,杨贵军,朱金山,等.基于小波变换与偏最小二乘耦合模型估测北方潮土有机质含量[J].光谱学与光谱分析,2014,34(7):1922－1926.[WANG Yancang,YANG Guijun,ZHU Jinshan,et al. Estimation of organic matter content of north fluvo-aquic soil based on the coupling model of wavelet transform and partial least squares[J]. Spectroscopy and Spectral Analysis,2014,34(7):1922－1926.]

[20] 蔡亮红,丁建丽.基于高光谱多尺度分解的土壤含水量反演[J].激光与光电子学进展,2018,55(1):013001[CAI Lianghong,DING Jianli. Inversion of soil moisture content based on hyperspectral multi-scale[J]. Laser & Optoelectronics Progress,2018,55(1):013001.]

[21] 唐梦迎,丁建丽,夏楠,等.干旱区典型绿洲土壤有机质含量分布特征及其影响因素[J].土壤学报,2017,54(3):759－766.[TANG Mengying,DING Jianli,XIA Nan,et al. Distribution of soil organic matter content and its affecting factors in oases typical of arid region[J]. Acta Pedologica Sinica,2017,54(3):759－766]

[22] 刘广崧,蒋能慧,张连第.土壤理化分析与剖面描述[M].北京:中国标准出版社,1996:166－167.[LIU Guangsong,JIANG Nenghui,ZHANG Liandi. Soil physical and chemical analysis and profile description[M]. Beijing: Standrds Press of China,1996:166－167.]

[23] 中国土壤学会农业化学专业委员会.土壤农业化学常规分析方法[M].北京:科学出版社,1989.[Agricultural chemical specialized committee of china soil society[M]. Conventional analytical method of soil agricultural chemistry[M]. Beijing: Science Press,1989.]

[24] 陈至坤,张菡洁,王玉田,等.基于小波变换的矿物油荧光光谱数据处理方法[J].激光杂志,2016,37(10):78－81.[CHEN Zhikun,ZHANG Hanjie,WANG Yutian,et al. Fluorescence spectral date of mineral oil processing based on wavelet transform[J]. Laser Journal,2016,37(10):78－81.]

[25] 刘燕德,欧阳爱国,应义斌.小波分析用于光谱信号处理及其在Matlab中的实现[J].传感技术学报,2006,19(3):821－823.[LIU Yande,OUYANG Aiguo,YING Yibin. Application of wavelet analysis in signal process using Matlab[J]. Chinese Journal of Sensors and Actuators,2006,19(3):821－823.]

[26] 沈润平,郭佳,张婧娴,等.基于随机森林的遥感干旱监测模型的构建[J].地球信息科学学报,2017,19(1):125－133.[SHEN Ruiping,Guo Jia,ZHANG Jingxian,et al. Construction of a drought monitoring model using the random forest based remote sensing[J]. Journal of Geo-information Science,2017,19(1):125－133.]

[27] VAUDOUR E,GILLIOT J M,BEL L,et al. Regional prediction of soil organic carbon content over temperate croplands using visible near-infrared airborne hyperspectral imagery and synchronous field spectral[J]. International Journal of Applied Earth Observation & Geoinformation,2016,49:24－3.

[28] SHI Z,WANG Q L,PENG J,et al. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations[J]. Science China Earth Sciences,2014,57(7):1671－1680.

[29] 高志海,白黎娜,王琫瑜,等.荒漠化土地土壤有机质含量的实测光谱估测[J].林业科学,2011,47(6):9－16.[GAO Zhihai,

BAI Lina, WANG Bingyu, et al. Estimation of soil organic matter content in desertified lands using measured soil spectral data[J]. Scientia Silvae Sinicae, 2011, 47(6): 9 – 16.]

[30] 李洪. 官厅水库消落带土壤有机质分布特征及其高光谱反演研究[D]. 北京:首都师范大学, 2014: 40 – 60. [LI Hong. Distribution characteristics of soil organic matter and its hyperspectral retrieval in the water-level-fluctuating zone of guanting reservoir [D]. Beijing: Capital Normal University, 2014: 40 – 60.]

Hyperspectral detection of soil organic matter content based on random forest algorithm

BAO Qing-ling[1,2], DING Jian-li[1,2], WANG Jing-zhe[1,2], CAI Liang-hong[1,2]

(1 *Key Laboratory of Wisdom City and Environmental Modeling Department of Education, Xinjiang University, Urumqi* 830046, *Xinjiang, China*; 2 *Key Laboratory of Oasis Ecology, Xinjiang University, Urumqi* 830046, *Xinjiang, China*)

Abstract: In order to explore how to retain the spectral information and accurately detect the soil organic matter content, this paper investigated the possibility of using spectral processing techniques such as wavelet decomposition and random forest method to estimate the soil organic matter content and analyze the spectral curves of different wavelet decomposition reconstruction spectra in different soil types using spectroscopy data. This study took the soil samples as the study objects which were collected in Weigan River Oasis of Kuqa County, a typical arid area oasis at north-central of the Tarim Basin in Xinjiang, China. The soil organic matter content of these samples was determined. The ASD Field Spec FR was used to measure the soil samples' spectrum, and the spectral data were preprocessed by wavelet decomposition and mathematical transformation. Discrete wavelet transform (DWT) has the function of multi-scale analysis, which can transform multi-scale wavelet decomposition of soil near infrared spectroscopy data to analyze the spectral curves of different wavelet decomposition reconstruction spectra in different organic matter content and different soil types. The correlation analysis was used to determine the maximum wavelet decomposition layer and filter sensitive bands. Finally, a multi-variant linear prediction model about soil organic matter content was established based on the optimal characteristic spectrum produced by combining grey correlation analysis, random forest method to analyze the significance of different wavelet decomposition characteristic spectra. The results showed as follows: (1) The spectral reflectance of each wavelet decomposed is decreased with the increase of organic matter content. At the same time, the spectral curve of cultivated soil and forest soil shows a more gradual change than that of the saline soil and desert soil. (2) The correlation between the decomposition spectrum of the wavelet transform and the soil organic matter content is decreased first and then increased with the increase of the decomposition layer. In the sixth layer, the characteristic spectral curve and the number of sensitive bands tend to be stable, which helps to determine this layer as the largest decomposition layer of wavelet transform. (3) Compared with the gray correlation analysis, the random forest model is in line with the expectation for screening the factors of wavelet decomposition at each layer, and it comes a list of descending order according to the impact on soil organic matter content as follows: $L3$-$(1/\mathrm{Lg}R)'$, $L4$-$(1/\mathrm{Lg}R)'$, $L6$-$(1/\mathrm{Lg}R)'$, $L5$-$(1/\mathrm{Lg}R)'$, $L2$-$(1/\mathrm{Lg}R)'$, $L0$-$1/\mathrm{Lg}R$, $L1$-$1/\mathrm{Lg}R$. (4) Combining all *SOM* estimation models for statistical analysis, the model based on *L-MC* has the highest accuracy. The research shows that the monitoring of soil spectral organic matter content based on machine learning classification method combined with wavelet decomposition can effectively reduce noise band interference and improve the classification prediction accuracy of feature bands. The random forest prediction classification model has significant advantages over the traditional linear prediction classification model, such as gray correlation analysis. The random forest model not only outperforms the grey correlation analysis in statistical results, but also shows better reliability and stability in predicting ability. The results could provide scientific reference and support for the study of soil nutrients in the arid zone and local precision agriculture.

Key words: hyperspectral; soil organic matter content; wavelet transform; random forest